We Claim:

Sub A1

1. A method for providing at least one pull service and at least one push service to a plurality of mobile users comprising the steps of:

reducing access latency for said at least one pull service running on at least one Web server by prefetching documents into a cache of at least one proxy gateway by using at least one factor relating to a frequency of access of said plurality of mobile users to said pull content of said pull service, an update cycle of said pull content and response delay for fetching said pull content from said at least one Web server to at least one proxy gateway, said at least one proxy gateway connected between said mobile user and said Web server; and

iteratively estimating a state of each of said plurality of mobile users for determining push content to be forwarded to said mobile user by said at least one push service running on said at least one Web server.

2. The method of claim 1 wherein said pull content is plurality of documents and said step of reducing access latency comprises the step of selecting a predetermined number of documents to be prefetched into said cache of said proxy gateway, wherein said predetermined number of documents have the greatest reduction in said access latency.

3. The method of claim 1 wherein said step of reducing access latency uses said factor of said frequency of access wherein frequently accessed documents are prioritized for being stored in a cache of a proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service.

4. The method of claim 1 wherein said step of reducing access latency uses said factor of said update cycle wherein said pull documents having a shorter update cycle are prioritized for being stored into a cache of a proxy gateway said proxy gateway being connected between said mobile user and said pull service and push service.

5. The method of claim 1 wherein said access latency uses said factor of said response delay wherein said pull documents having a longer response delay are prioritized for being stored in a cache of a proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service.

6. The method of claim 1 wherein said step of reducing access latency comprises the step of selecting a predetermined number of documents to be prefetched into cache of a proxy gateway, and said step of selecting a predetermined number of documents uses said factors of: said frequency of access, said update cycle and said response delay, wherein said frequently accessed pull documents having a shorter update cycle and a longer response delay are prioritized for being prefetched in said cache of said proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service.

7. The method of claim 1 wherein said step of iteratively estimating a state of said mobile user is determined from tracking data of said plurality of mobile users and geo-location measurement and behavior observation data.

8. The method of claim 7 further comprising the step of:

caching mobility and behavior-related push content into a cache of said proxy gateway connected between said plurality of mobile users and said at least one Web server.

9. The method of claim 8 wherein each said state of one of said mobile users is determined from at least one of the following factors: location of said one of said plurality of mobile users, direction of said one of said plurality of mobile users, speed of said one of said plurality of mobile users, and behavior of said one of plurality of mobile users.

10. A system for providing at least one pull service and at least one push service to a plurality of mobile users comprising:

means for reducing access latency for said at least one pull service running on at least one Web server by prefetching documents into a cache of a proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service by using at least one factor relating to a frequency of access of said plurality of mobile users to said pull content of said pull service, an update cycle of said pull content and response delay for fetching said pull content from said at least one Web server to said at least one proxy gateway said at least one proxy gateway connected between said mobile user and said Web server; and

means for iteratively estimating a state of each of said plurality of mobile users for determining push content to be forwarded to said mobile user by said at least one push service running on said at least one Web server.

11. The system of claim 10 wherein said pull content is a plurality of documents and said means for reducing access latency comprises the step of:

selecting a predetermined number of documents to be prefetched into a cache of a proxy gateway; and

selecting a predetermined number of documents uses said factors of: said frequency of access, said update cycle and said response delay, wherein said pull documents having a higher frequency of access, a shorter update cycle and a longer response delay are prioritized for being prefetched in said cache of said proxy gateway.

12. The system of claim 10 wherein said means for iteratively estimating a state of said mobile user uses tracking data of said plurality of mobile users and geo-location measurement and behavior observation data.

13. The system of claim 12 further comprising:

means for controlling of caching mobility and behavior-related push content into a cache of said proxy gateway connected between said plurality of mobile users and said at least one Web server.

14. The system of claim 13 wherein each said state of one of said mobile users is determined from at least one of the following factors: location of said one of said plurality of mobile users, direction of said one of said plurality of mobile users, speed of said one of said plurality of mobile users, and behavior of said one of plurality of mobile users.

15. In a system comprising a proxy gateway connected by a first network to a plurality of mobile users and by a second network to at least one Web server, said proxy gateway comprising a cache for storing pull content received from said at least one Web server of a pull service, a method comprising the steps of:

storing data that is indicative of a request for said pull content from at least one of said plurality of mobile users and data indicative of interactions between said cache and said Web server;

determining access probability of access to said pull content from said stored data;

determining an average hit rate for said pull content from said stored data;

determining said average response delay for said pull content from said stored data;

determining average wired network access latency for said pull content from said access probability, said average hit rate and said average response delay;

storing said pull content in said cache based on said determined average wired network access latency when there is no said pull content in said cache or said pull content has expired,

wherein said pull content having a greater average wired network access latency is prioritized for being stored in said cache.

16. The method of claim 15 wherein said pull content is a plurality of n documents, n=1, 2 ... N, wherein N is the total number of documents, and said stored data comprises:

an average rate of access to document n, R_n; a size of said document n, s_n; an average time delay imposed by said second network, ΔT_n; and an update cycle of said document n, μ_n.

17. The method of claim 16 wherein said access probability is determined by:

$\gamma_n = R_n / R$

wherein R is the total rate of access traffic on said second network.

18. The method of claim 17 wherein said average hit rate for document n, h_n is determined by:

$$h_n = 1 - \frac{g_n}{R_n \mu_n}, \quad n=1, 2, \ldots, N,$$

in which:

g_n is the probability that there is at least one request to document n during a given update cycle, μ_n, given by:

$$g_n = 1 - e^{-R_n \mu_n}, \quad n = 1, 2, \ldots, N, \qquad (3)$$

and

$R_n\mu_n$ is the expected number of accesses to document n in an update cycle of document n.

19. The method of claim 18 wherein average wired-network-access latency when there is no said pull content in said cache or said pull content has expired is determined from

$$\eta_n = \gamma_n (1-h_n) \Delta T_n, \quad n = 1, \ldots, N,$$

20. The method of claim 19 wherein said pull content is prioritized by the steps of:

sorting said plurality of N documents in descending order with the document having the greatest average wired network access latency when there is no said pull content in said cache or said pull content has expired labeled as η_1, and the document having the least average wired-network-access latency when there is no said Web content in said cache or said Web content has expired labeled as η_N; and

determining a number of documents to be stored in said cache, r, by considering at least one constraint selected from the group consisting of spare cache capacity, spare transmission bandwidth on said second network and desired hit probability.

21. The method of claim 20 wherein said constraint of said spare cache capacity, ΔC, is given by:

$$\sum_{n=1}^{r} s_n (1-h_n) \leq \Delta C,$$

wherein $\Delta C \approx C - \sum_{n=1}^{N} s_n h_n$, C is given capacity of the cache, s_n is the size of the document n and h_n is the average hit rate for document n.

22. The method of claim 20 wherein said constraint of said spare transmission bandwidth, ΔB, is given by

$$\sum_{n=1}^{r} (1-g_n) \frac{S_n}{\mu_n} \leq \Delta B,$$

wherein $\Delta B \approx B - \sum_{n=1}^{N} g_n \frac{S_n}{\mu_n}$, B is given bandwidth, g_n is the probability that there is at least one request to document n during a given update cycle μ_n and s_n is the size of the document.

23. The method of claim 20 wherein said constraint of said desired minimum hit probability, ΔH, is given by:

$$\sum_{n=1}^{r} \gamma_n \ (1\text{-}h_n) \geq \Delta H$$

wherein $\Delta H \approx H - \sum_{n=1}^{N} \gamma_n \ h_n$ H is given hit probability, γ_n is an access to document n and h_n is an average hit rate for document n.

24. The method of claim 16 further comprising the step of:

updating said stored pull content in said cache based on said update cycle of document n, μ_n.

25. In a system comprising a proxy gateway connected by a first network to a plurality of mobile users and by a second network to at least one Web server, a method comprising the steps of:

measuring each of said mobile users current geo-location position and behavior;

computing a first probability that said measured current geo-location position and behavior is an actual position and behavior of each of said mobile users;

determining a state sequence estimation variable for each of said mobile users by iteration over time from a second probability that each of said mobile users transit in a geo-location and behavior sequence;

determining a current state for each of said mobile users from said state sequence estimation; and

pushing push content related to said current state to each of said mobile users.

26. The method of claim 25 wherein said first probability is given by

$$P_r \{Y_t \mid X\}$$

wherein Y_t is said measured current geo-location position and behavior and X is said actual position and behavior.

27. The method of claim 26 wherein said state sequence estimation variable is determined by

$$\alpha_t(m) = \sum_{m'=0}^{M-1} \alpha_{t-1}(m') \ p_{m'm} \sum_{x} \Pr\{x \mid m\} \Pr\{Y_t \mid x\}$$

wherein $p_{m'm}$ is the state transition probability of one of said plurality of mobile users, $\Pr\{x \mid m\}$ is the probability that said one of said plurality of mobile users locates at position and behavior as x when it is in state m at time t, and $\Pr\{Y_t \mid x\}$ is the probability that said

measured geo-location is Y_t when said one of said plurality of mobile users position and behavior is x at time t.

28. The method of claim 27 wherein said current state is determined by $z = \arg\max_{m} \{\alpha_t(m) \mid m = 1,2,\ldots,M-1\}$.

29. The method of claim 28 wherein said proxy gateway comprising a cache for storing push content received from said at least one Web server and said push content is stored in said cache based on said current state.